Boston Theater Marketing – An Algorithm for Advertising Venue Choices

Troy Stedman

JUNE 24, 2020

Table of Contents

A. Abstract

An algorithm for advertising a business to local venues is developed. This algorithm uses free location data from the Foursquare API. Using the “rating”, “likes”, and “distance” attributes for a venue as well as the “nextvenues” endpoint query, a model is constructed that will sort the top 100 venues near the location of the business to be advertised. This algorithm could potentially be used by marketing agencies.

1. Introduction/Business Problem

For any new business, developing a customer base is crucial. A proven strategy to attract customers is through advertising. However, efficient advertising can be difficult, especially for a new business. Therefore, when opening a business in a physical location, it is important to know which local venues would be optimal to advertise this new business. The purpose of this project is to develop an algorithm to determine which local venues would be best to advertise a business in a set location. Marketing agencies and advertisers could potentially use this model for their services. This algorithm is applied to the fictitious example below, although the overall model can be applied in more general cases.

To illustrate the model, consider the following example. A patron of classical arts wants to open a theater near Boston City Hall. Of course, as this will be a new theater, the patron needs to attract customers to develop a customer base. To do so, he contacts a marketing agency to find out which places in Boston would be best to advertise the new theater. These places should constitute an optimal set of venues for both physical and digital advertising. To carry out the project, the marketing agency will use the Foursquare API to collect information about venues in Boston and construct a system to rank which venues would be best for advertising the new theater. Foursquare is a data and intelligence company that provides location data for venues around the world. For example, one could use Foursquare as a simple search engine or for a more detailed description of venues. Developers may use the Foursquare API to freely gather limited location data or have access to more location data for a fee. Unfortunately, due to lack of funding, the data gathered from the Foursquare API will only be the free data accessible to developers. The marketing agency hopes that this project would be able to be further improved and models further generalized for a wider audience of advertisers.

2. Data: Collection and Processing

The marketing agency will come up with a ranking system to determine the best venues in Boston for advertising the theater near Boston City Hall. The data to be used in this ranking system will come exclusively from the Foursquare API.

First, the marketing agency will gather data for the top 100 venues around Boston City Hall (using latitude and longitude geographic coordinates) based on the Foursquare API “explore” endpoint query. The explore endpoint shows the top venues ranked by Foursquare around a location. Location data for these venues are then available. Boston City Hall has geographic coordinates (42.360100, -71.058900) for latitude and longitude, respectively. An example of the available location data results shown by this explore endpoint for Boston City Hall is shown in Table 1 for the top 5 venues from this search.

Table 1. *The top 5 venues around Boston City Hall according to the Foursquare API explore endpoint. The name, category (type of venue), and latitude and longitude location for each venue are given. In addition, the distance from Boston City Hall in meters for each venue are shown.*

Name	Category	Latitude	Longitude	Distance (meters)
Boston Athenaeum	Library	42.35748	-71.0618	378
North End Park	Park	42.36249	-71.0565	332
Faneuil Hall Marketplace	Historic Site	42.35998	-71.0564	205
haley.henry	Restaurant	42.35757	-71.0595	285
The Rose Kennedy Greenway - Mothers Walk	Park	42.36264	-71.0564	349

2.1 "Next Venues" Data

After using the explore endpoint to find the top 100 venues around Boston City Hall, the marketing agency will use a methodology applied to these venues to determine the most appropriate venues for advertising. The methodology is based on the idea that venues from which people then went to theaters would be good sites for theater advertising. In other words, if people went to a theater after visiting a particular venue, then that venue is a good place to advertise for a theater. This is a reasonable assumption since people who go to theaters are more likely to go to a theater again than people who don't go to theaters. With this observation, a filtering process will be used based on the "nextvenues" endpoint from the Foursquare API. This endpoint shows the top 5 venues people visited from a given venue. Table 2 shows a sample of the nextvenues results for the top 5 venues from the explore endpoint for Boston City Hall shown in Table 1. The data that will be used from the nextvenues endpoint will be the category for each of the 5 next venues. Missing data (indicated by np.nan) are ignored in the model.

Table 2. *The 5 next venues and their category that people visited from the venues listed in Table 1. Note some missing data occur.*

Name	Next Venue 1/ Category	Next Venue 2/ Category	Next Venue 3/ Category	Next Venue 4/ Category	Next Venue 5/ Category
Boston Athenaeum	Boston Common/ Park	Scollay Square/ American Restaurant	Emmet's Irish Pub/ Irish Pub	Faneuil Hall Marketplace/ Historic Site	21st Amendment/ Pub
North End Park	Mike's Pastry/ Pastry Shop	Paul Revere House/ Historic Site	Quincy Market/ Historic Site	Faneuil Hall Marketplace/ Historic Site	Neptune Oyster/ Seafood Restaurant
Faneuil Hall Marketplace	Quincy Market/ Historic Site	Cheers/ Bar	Mike's Pastry/ Pastry Shop	Newbury Comics/ Record Shop	Paul Revere House/ Historic Site

haley.henry	Yvonne's/ New American Restaurant	JM Curley/Gastropub	np.nan	np.nan	np.nan
The Rose Kennedy Greenway - Mothers Walk	Paul Revere House/ Historic Site	Faneuil Hall Marketplace/ Historic Site	Boston Public Market/ Market	Mike's Pastry/ Pastry Shop	Neptune Oyster/ Seafood Restaurant

In addition to the nextvenues category data for the top 100 venues near Boston City Hall, specific attributes about each of these venues will be used: “rating”, “likes”, and “distance” from Boston City Hall. The rating attribute is the average rating given to a particular venue by Foursquare users, the likes attribute is the total number of Foursquare users that liked a particular venue, and the distance attribute in this case is just the distance in meters of a particular venue from Boston City Hall. Thus, the relevant location data to be used in the ranking system are given below.

Foursquare Location Data to be Used:

- Top 100 venues from the explore endpoint for Boston City Hall.
- Rating, likes, and distance (from Boston City Hall) attributes for each of these top 100 venues.
- The category of each venue in the nextvenues endpoint for each of these top 100 venues.

3. Methodology

Now, each venue has a uniquely named “category” attribute that describes the type of venue. Foursquare gives a list of all possible categories. There are 10 main categories: “Arts & Entertainment”, “College & University”, “Event”, “Food”, “Nightlife Spot”, “Outdoors & Recreation”, “Professional & Other Places”, “Residence”, “Shop & Service”, “Travel & Transport”. For each main category, there is a tree of subcategories (nodes). These trees are main category trees. Each node in these trees corresponds to a possible venue category and may have a parent node and children nodes. Therefore, each possible category corresponds to a specific node in the collection of main category trees. We connect all the main category trees by giving the main categories a common parent node called “Category” that will serve as the starting node for the entire tree of categories. The unique path for each category in its corresponding main category tree can be found from following the parent nodes starting at the node for the particular category. A list of these parent nodes can be constructed to give the path for the category. For example, the “Theater” category has “Performing Arts Venue” as its parent node, which has Arts & Entertainment as its parent node (this node is also a main category). Therefore, the path for the Theater category can be written as [Category, Arts & Entertainment, Performing Arts Venue, Theater].

Given a path list, each category in the path has a corresponding *depth* which is simply its position within the path. For example, for the Theater path above, Category has a depth of 0 (which is always the case), Arts & Entertainment has a depth of 1 since it is the first category in this path. Similarly, Performing Arts Venue has a depth of 2 and Theater has a depth of 3. Note that these depth values are actually

attributes for each category since every path that contains these categories will have these categories at the same depths described above. Therefore, the depth of a category is unique. Figure 1 shows the category tree, with depths for each node, and an illustration of the Theater and "Dance Studio" paths.

Theater Path: [Category, Arts & Entertainment, Performing Arts Venue, Theater]

Dance Studio Path: [Category, Arts & Entertainment, Performing Arts Venue, Dance Studio]

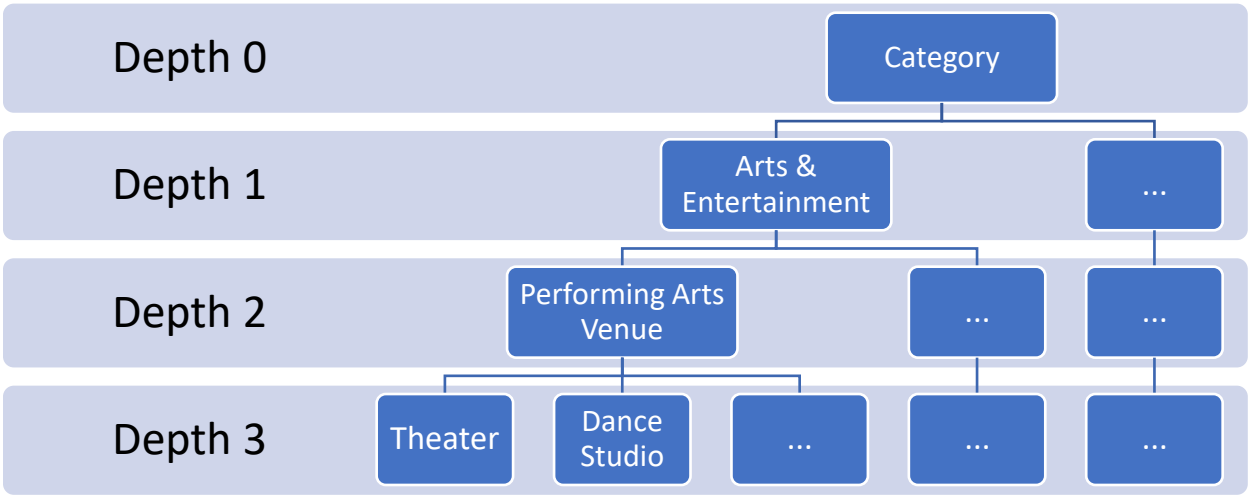


Figure 1. *A tree of all categories and the depths of each category node. Ellipses indicate extra branches and categories that may even go to higher depths than shown in the figure.*

3.1 Similarity Index

Given two categories, we would like to know how related they are using their unique paths. We can compare these paths by determining if they have any nodes in common. The common node with the largest depth determines the similarity of the two categories. We define the similarity index to be the maximum depth of all common nodes for two categories (determined by their paths). The maximum a similarity index can be is then 3 (the depth of Theater in its own path) and the minimum a similarity index can be is 0 (the depth of Category for any path). As an example, the path for the Dance Studio category is [Category, Arts & Entertainment, Performing Arts Venue, Dance Studio]. Comparing this path to the Theater category path, we see that the two categories have a similarity index of 2 through the Performing Arts Venue category of depth 2. On the other hand, if we compare the Theater category path to the "American Restaurant" path [Category, Food, American Restaurant], we see that the American Restaurant and Theater categories have a similarity index of 0. Naturally, the similarity index therefore is a measure of how similar two categories are. An illustration of the similarity index for two categories is shown in Table 3.

Table 3. *The paths of the Theater and Dance Studio categories with depths for each subcategory. The similarity index is 2 since the maximum depth of common subcategories has depth 2.*

	Depth 0	Depth 1	Depth 2	Depth 3	Similarity Index
Theater	Category	Arts & Entertainment	Performing Arts Venue	Theater	2
Dance Studio	Category	Arts & Entertainment	Performing Arts Venue	Dance Studio	

For each venue in the nextvenues list, we calculate the similarity index of their respective category to the Theater category. The similarity indices calculated for the nextvenues endpoint allow us to measure the likelihood a Theater category would be visited. For a given venue, if the venues in the nextvenues list have similarity indices close to 3 for the Theater category, then we would expect the given venue to be a good candidate for advertising as the venues visited from the given venue are similar to Theater venues. Two quantities based on the similarity indices for nextvenues searches will be used. First, the maximum similarity index of the nextvenues list of a given venue will be used. This quantity will be called the "Max Next Venue Similarity Index". Second, the average of the similarity indices of the nextvenues list of a given venue will also be used. This quantity will be called the "Average Next Venue Similarity Index". Any missing nextvenues entries will be ignored for the max next venue similarity index and the average next venue similarity index.

The classification will break down venues based on the max next venue similarity index and average next venue similarity index. The max next venue similarity index is given the highest priority and the average next venue similarity index is given the next highest priority. In other words, venues will be sorted first based on the max next venue similarity index. Then, each venue with the same Max Next Venue Similarity Index will be sorted based on the Average Next Venue Similarity Index. The result of this sorting procedure is equivalent to the sorting done on a Pandas DataFrame by pandas.DataFrame.sort_valules(by=[Max Next Venue Similarity Index, Average Next Venue Similarity Index]). The sorting priority is then given by

Max Next Venue Similarity Index → Average Next Venue Similarity Index

and the venues are grouped according to this priority setting to give the groups described by the pair (Max Next Venue Similarity Index, Average Next Venue Similarity Index). For example, if a venue has Max Next Venue Similarity Index = 2 and Average Next Venue Similarity Index = 1.2, then this venue falls in the group (2, 1.2). Table 4 shows next venue similarity index data for select venues and their groups. Again note that missing data are ignored.

Table 4. *Next venues similarity index data for select venues. The venues are grouped according to the sorting procedure described above to give groups described by the pair (Max Next Venue Similarity Index, Average Next Venue Similarity Index). Groups for these select venues are color coded: green is for the highest ranking group (1, 0.6), blue is for the next highest ranking group (1, 0.4), yellow is for the next highest ranking group (1, 0.2), and red is for the lowest ranking group (0, 0).*

Name	Next Venue 1 Similarity Index	Next Venue 2 Similarity Index	Next Venue 3 Similarity Index	Next Venue 4 Similarity Index	Next Venue 5 Similarity Index	Max Similarity Index	Next Venues Average Similarity Index
Boston Athenaeum	0	0	0	1	0	1	0.2
North End Park	0	1	1	1	0	1	0.6
Faneuil Hall Marketplace	1	0	0	0	1	1	0.4
haley.henry	0	0	np.nan	np.nan	np.nan	0	0
The Rose Kennedy Greenway - Mothers Walk	1	1	0	0	0	1	0.4

4.2 Minkowski Length

After this sorting procedure is applied, further sorting may be needed since some venues may have the same Max Next Venue Similarity Index and Average Next Venue Similarity Index. Since the nextvenues data have already been used for this sorting procedure, we use one last quantity for the final sorting. It would be reasonable that this quantity should capture attributes related to the popularity and quality of a venue. The Foursquare API venue attributes to be included in the sorting model are the rating, likes, and distance from Boston City Hall. These venue attributes require Premium API calls and can be found through a "details" endpoint call. We want to find those venues with a large rating, large likes, and small distance since we would like to advertise to venues close to Boston City Hall. To this end, we will calculate an unweighted Minkowski length based on these attributes. First, we normalize these attributes by their max values within each (Max Next Venue Similarity Index, Average Next Venue Similarity Index) group. For the distance, we use 1 minus the normalized distance in calculating the Minkowski length. Then, we calculate the unweighted Minkowski length according to

$$\sqrt{(Normalized\ Rating)^2 + (Normalized\ Likes)^2 + (1 - Normalized\ Distance)^2}.$$

As a remark, this formula gives a maximum possible Minkowski length of $\sqrt{3} \approx 1.732051$ when a venue has the max rating, max likes, and a distance of 0 in its group and a minimum possible Minkowski length of 0 when a venue has a rating of 0, 0 likes, and the max distance of its group. Also, if a venue has the max rating, max likes, and max distance of its group, the Minkowski length will be $\sqrt{2} \approx 1.414214$. Table 5 shows the Minkowski lengths for a set of hypothetical venues and groups.

Table 5. *Minkowski lengths for hypothetical venues "A", "B", "C", "D", and "E" under groups (1, 0.8) and (1, 0.5) color coded by green and yellow, respectively. The max rating, likes, and distance for each group are color coded as well.*

Group: (Max Similarity Index, Average Similarity Index)	Venue	Rating	Likes	Distance	Minkowski Length
(1, 0.8)	A	9	157	241	1.414214
	B	8.5	80	197	1.088556
(1, 0.5)	C	9.1	625	237	1.415944
	D	8.7	449	303	1.179504
	E	9.3	595	172	1.446798

4. Results

The final sorting procedure uses a priority sorting according to

Max Next Venue Similarity Index → Average Next Venue Similarity Index → Minkowski Length.

This sorting procedure was applied to the top 100 venues near Boston City Hall and the top 10 ranked venues based on this sorting procedure are shown below in Table 6.

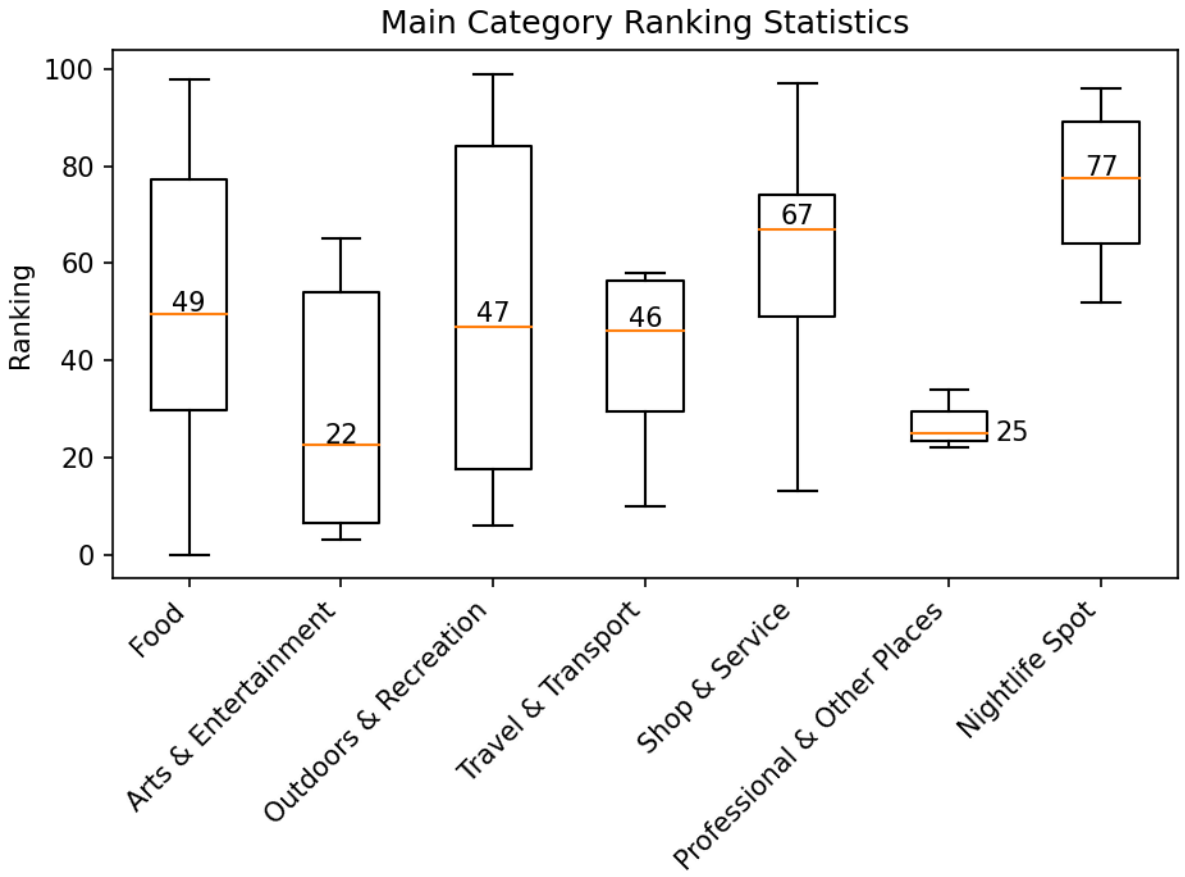
Table 6. *The top 10 venues of the final rankings based on the sorting procedure developed.*

Rank	Name	Main Category	Average Next Venues Similarity Index	Minkowski Length
1	Picco	Food	3	1.2
2	sweetgreen	Food	2	0.5
3	Tatte Bakery & Cafe	Food	1	1
4	New England Aquarium	Arts & Entertainment	1	0.8
5	Aquarium Seal Tank	Arts & Entertainment	1	0.8
6	Museum of Science	Arts & Entertainment	1	0.6
7	North End Park	Outdoors & Recreation	1	0.6
8	The Freedom Trail	Arts & Entertainment	1	0.6
9	Boston Harborwalk	Outdoors & Recreation	1	0.6
10	Flour Bakery & Cafe	Food	1	0.6

4.1 Exploratory Analysis of Results

The top 10 venues according to the sorting procedure are, respectively: Picco, sweetgreen, Tatte Bakey & Cafe, New England Aquarium, Aquarium Seal Tank, Museum of Science, North End Park, The Freedom Trail, Boston Harborwalk, Flour Bakery & Cafe.

Only 1 venue has a Max Next Venue Similarity Index of 3 (Picco) and only 1 venue has a Max Next Venue Similarity Index of 2 (sweetgreen). All other venues have a Max Next Venue Similarity Index less than 2. The top 3 venues are in the Food main category, while the 4th ranking venue is in the Arts & Entertainment main category. Since the main categories of the top 100 venues near Boston City Hall are not a part of the sorting model, the statistics of these main categories with regards to the ranking may be of interest. To get a breakdown of these statistics, a boxplot and statistics table are provided in Figure 2.



Main Category	Count	Mean	Standard Deviation	Min	1st Quartile	Median	3rd Quartile	Max
Food	48	51.6875	28.14866	0	29	49	77	98
Arts & Entertainment	12	30.08333	24.97438	3	6	22	54	65
Outdoors & Recreation	20	51	32.97846	6	17	47	84	99

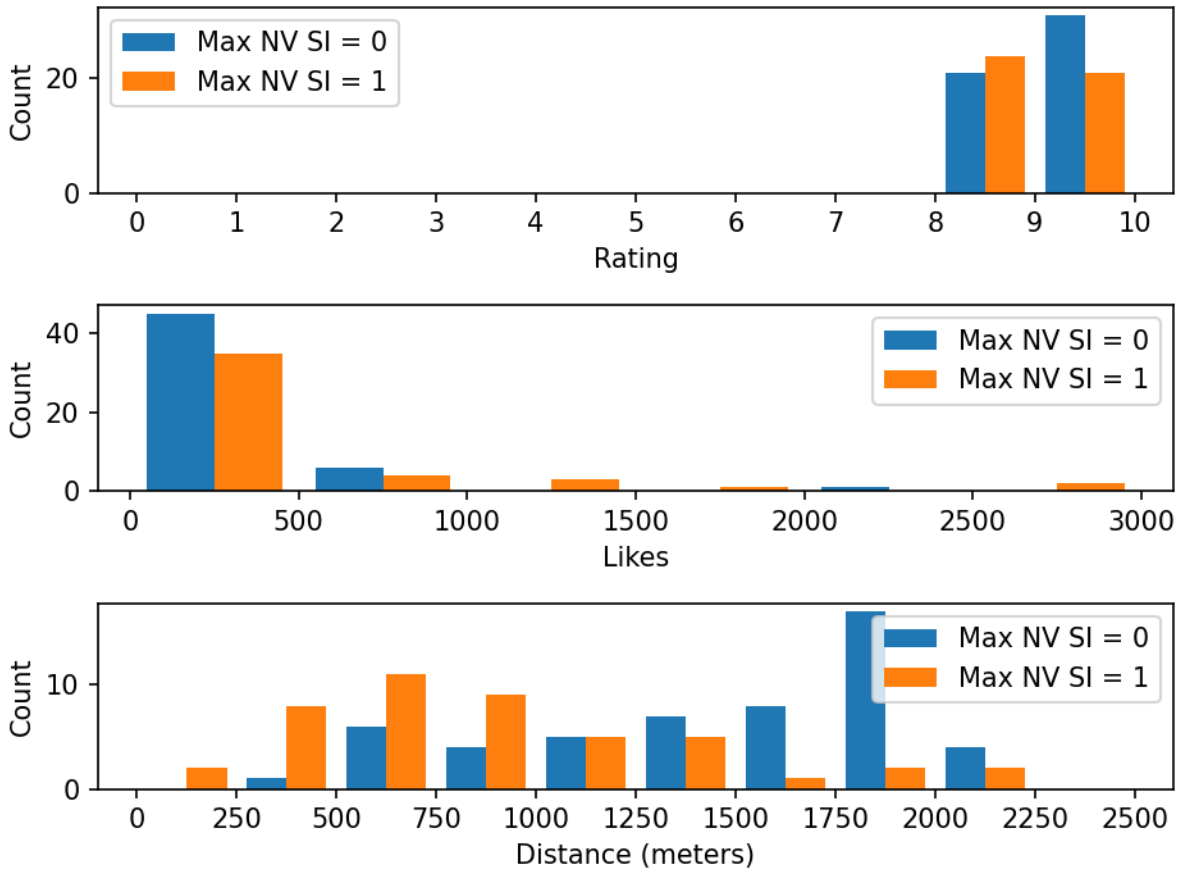
Travel & Transport	4	40	22.33085	10	29	46	56	58
Shop & Service	9	60.44444	25.76874	13	49	67	74	97
Professional & Other Places	3	27	6.244998	22	23	25	29	34
Nightlife Spot	4	75.75	19.6702	52	64	77	89	96

Figure 2. *A boxplot of the ranking statistics for the main categories of the ranked top 100 venues near Boston City Hall. The medians are also shown on the graph. A table of statistics for these main categories is also shown.*

From Figure 2, the Food category has the most occurrences with 48 venues while the Professional & Other Places category has the least occurrences with 3 venues. Professional & Other Places also has the lowest mean rank of 27, with a small standard deviation of 6.244998. Nightlife Spot has the highest mean rank of 75.75, suggesting that (unsurprisingly) few people go to theaters after visiting a Nightlife Spot. Interestingly, there is generally a large spread between the Min and Max ranks for each category, except for the Professional & Other Places category. These statistics suggest that the ranking is more specific to the venues around Boston City Hall rather than describing a general pattern at other locations.

As there are only 2 venues with a Max Next Venue Similarity Index greater than 1, these venues are seemingly particular to this specific application of the sorting model. A further exploratory analysis will be applied to the remaining venues which have a Max Next Venue Similarity Index less than or equal to 1. For these remaining venues, let's take a look at their distributions across the rating, likes, and distance attributes. Figure 3 shows a collection of these distributions and their correlations and p-values with the Max Next Venue Similarity Index.

4.2 Histograms of Main Categories



	Max Next Venue Similarity Index (Pearson Correlation, P-Value)
Rating	(-0.0649565, 0.527307)
Likes	(0.2227224, 0.0283257)
Distance	(-0.5199544, 0.0)

Figure 3. *Histograms for the rating, likes, and distance attributes for the top 100 venues that have a Max Next Venue Similarity Index of 1 or 0. In the histograms, Max Next Venue Similarity Index is abbreviated as "Max NV SI". A table of Pearson correlations and p-values for the rating, likes, and distance attributes against the Max Next Venue Similarity Index for these data is also shown. The p-value of 0.0 for the distance attribute is simply due to rounding not including enough decimals.*

Judging by the histograms in Figure 3, we might surmise there is a negative correlation between likes and distance and the Max Next Venue Similarity Index, while there is likely no correlation between rating and the Max Next Venue Similarity Index. The Pearson correlations shown in Figure 3 show that there are no strong correlations between ratings, likes, and distance and the Max Next Venue Similarity Index. The largest correlation in magnitude was between the distance and the Max Next Venue Similarity Index, with a correlation of -0.5199544 and p-value of 0.0 (too small for the decimal rounding), indicating a statistically significant moderate negative correlation, while the smallest correlation in

magnitude was between the rating and the Max Next Venue Similarity Index, with a correlation of -0.0649565 and p-value of 0.527307, indicating a statistically insignificant weak negative correlation. These statistical results for the venue ratings make sense as the rating of a venue does not depend on whether or not visitors will go to a theater next. Put another way, the rating of a venue only indicates the perception of its quality by its visitors and not any relation to the next venues visited. The statistical results for the venue distances indicate that venues that are closer to Boston City Hall are generally preferred for theater goers, maybe because of the density of entertainment venues around Boston City Hall or some other factor. The correlation and p-value between likes and the Max Next Venue Similarity Index are respectively (0.2227224, 0.0283257), showing that the weak positive correlation is statistically significant. This positive correlation is too weak to be conclusive.

In general, the statistics of the top 100 ranked venues for an advertised venue depend on the advertised venue, as well as the top 100 venues themselves and the location of interest. Through the exploratory analysis and statistics of the top 100 ranked venues for theater advertising near Boston City Hall, we see these specific trends appear.

5. Discussion

The optimization of advertising for businesses was explored using a proposed sorting model. This model can be applied to venues at any location that has Foursquare API location data. As an example of how the model can be used, the ranking of the top 100 venues for advertising a theater near Boston City Hall was explored. The ranking of the top 100 venues for this advertising based on the sorting model were found. In addition, an exploratory analysis of these ranking results was performed. The results, including the ranking of the top 100 venues, are dependent on the advertising venue, the top 100 venues themselves, and the location of interest. In other words, the model should not be expected to produce general ranking patterns.

This sorting model could be of value to advertisers and marketing agencies as it is easy to implement and can be generalized further.

6. Conclusion

A sorting algorithm was used to construct a model that would rank the top 100 venues near Boston City Hall that would be the most appropriate to advertise a new theater. This model used location data from the Foursquare API. For each venue, the sorting mechanism first ranked results using the Max Next Venue Similarity Index and the Average Next Venue Similarity Index calculated from the nextvenues categories. The ratings, likes, and distance from Boston City Hall were used to calculate the Minkowski length for each venue that was used for the last sorting step. The output of the model would be a complete ranking of the top 100 venues based on the sorting model priority

Max Next Venue Similarity Index → Average Next Venue Similarity Index → Minkowski Length.

The sorting model was demonstrated here for advertising of a theater near Boston City Hall but can obviously be applied to other types of venues in other locations. The dependence of the model on the type of venue considered and the location should be considered. In general, the rankings produced for a specific type of venue and location might be more appropriate for this specific type of venue and location. This is because one type of venue might be more prevalent or popular than another, while also being dependent on the location. It is therefore important to recognize the particularities of the type of venue and location when applying the sorting model. To try and capture a broader more general behavior and to minimize any bias through location, the model can be applied to any area with Foursquare location data. If enough datasets are gathered for a set of locations, a predictive regression model can be setup to determine the rankings of venues in a city for a specific category. This could be useful if one does not have access to data like the nextvenues location data from the Foursquare API.

In addition to the different ways the sorting model can be applied, the model could be further improved by incorporating other venue features that might be of relevance, like the regular number of visitors or average age of the visitors. For further advertising projects or other business inquiries, the marketing agency should explore other features to include in the model and ways to improve accuracy based on advertising data. The accuracy of the sorting model should be investigated based on future data gathered. A feedback loop with this accuracy can then be setup to further optimize the model.